Artificial Intelligence Based Predictive Methods To Detect Cyber Security Attacks

Synopsis For the Doctoral Degree in Computer Science and Engineering

Submitted
By

Shivaji Digambarrao Jadhav

Under the Guidance of

Dr. B. R. Bombade
Associate Professor
Department of Computer Science and Engineering
Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded

Submitted
to
Research and Recognition Committee

SWAMI RAMANAND TEERTH MARATHWADA UNIVERSITY
NANDED – 431606, Maharashtra State, India
(Academic Year 2020-2021)

Contents

1. Introduction:----------3
2. Literature Review ----------4
3. Research Methodology: ----------7
3.1 Problem Statement:----------7
3.2 Methodology:----------8
4. Objective:---------- 10
5. Expected Outcomes: ---------- 10
References---------- 11

1. Introduction:

With the evolution of semiconductor-based systems from the early 19th Century to the present 21st century world, computer systems evolved a lot and boosted humankind efforts to discover novel methods of interconnected networks and supportive input-output, storage, and other devices. It changed the way humans computed and communicated in the early days. At early stages of development, the complexity of modern digital systems at architecture level made it complex to operate and manage. With the time, computer systems evolved drastically, and it became more sophisticated tool with highly complex but easy to operate and manage.

As the computer system starts to grow more, it also helped to grow the early computer system worms of 1980s to present day cyberattack scenarios. World has faced lot of cybercrimes, cyber atrocities, and cyber economic crimes and many more. Be it individual, private or government organization, all of them some time either became victim or heard about cybercrimes happed around them. When attacker enters in the system till the complete attack objectives are achieved, the attacker accesses several devices and services available on the system. These actions create a digital footprint and many a times, it is not possible for humans to analyze all such footprints e.g., Event logs before, during and after Cybercrime may be some time in Terabyte size. In real attack scenarios, where number of attached systems are highly large and the network structure is sufficiently complex, the footprints available like logs generated over the network and the transactions happening per sections creates millions of records. To analyze such events, there are open source and commercial software's which are used by the cyber security experts. But such software tools are used post attack analysis scenarios.

Finding Correlation: Correlate log data from different sources in real time to get a complete picture of the attack is highly needed. Research is needed to be carried out to find out the out-of-the-box correlation rules, to spot known attacks and create custom rules for newer threats with prediction of events by analyzing real time logs and post event logs (both). Advance research is needed to be done on all sort of event logs for creating powerful correlation engine that detects attack patterns instantly. It should identify every log trail of various incidents (as they span out over multiple devices in your system and over the network), and must alerts you of any suspicious attack scenario.

Basically log analysis is a time consuming process and till date no solution is available to predict and analyze the logs in real time scenarios and also in the post analysis scenarios. Research is needed to be done to solve the issues and problems in such Cyber Security Incident scenarios. Also, it is found that the cyber attacker tries to delete all the digital footprints available on the system and network devices post Cybercrime, hence it is a challenge to researchers to find out the Cybercrime and its perpetrator in real time.

In Cyberattack scenarios, breach may occur due to various reasons like unattended open port, open gateways, unsecured domains, operating without SSL certificates, old hardware and software's, absence of hardening of network devices, absence of two factor authentications, relying on unsecured third-party services, allowing use of putty commands from back-doors and many more. These are few of the reasons which leads to cyberattacks. Hence, it is expected to find out a novel method which can predict cyber security attack scenario in active or passive methods. One can use Artificial Intelligence (AI) based approach to predict such attacks where system can be trained to identify the cyber security attacks by analyzing various system and network generated logs, authentication process, access to third-party services policy, previous case studies to create a system learning curve and other such things.

2. Literature Review

Andre Brandao et al paper provides details on the aspect of developing system for Network intrusion detection through network logs [1]. It states that, Information security remains an unsolved challenge for organizations, as every day brings new and sophisticated cyber-attacks happens. The implementation of system to monitor the logs and, in case of any anomalous traffic, to detect and hopefully prevent any security incident but it is least available as these systems generate a huge number of logs per second to be handled, leading to the need of automated ways to identify the cyber-attacks. In this paper author propose an effective Log-based Intrusion Detection System (LIDS), to predict an attack or not, based on carefully selected features where the logs from various sources are aggregated into one dashboard and the most discriminative features are first determined. For the attack prediction, a few machine learning techniques were comparatively tested, with the Decision tree being the winner. The proposed system is illustrated with the largest publicly available labelled log file dataset KDD Cup 1999. It is observed that. here input is a very old dataset and logs are from limited sources.

In Július Baráth et al paper, the author first does the collection and then analysis of event logs which allows detection and debugging of operating system and application configuration errors [2]. Author states that appropriate selection of event logs allows to detect cyber-attacks and prevent potential damage. In the article, author is more focused on the selection and optimization of event logs for the Microsoft Windows operating system. Author has experimentally verified the structure and number of produced logs and we proposed their optimization. Author is limited to only Windows systems based analysis.

Yuan et al paper provides details on, Event logs, commonly available in modern mechatronic systems, which contain rich information on the operating status and working conditions of the system [3]. This article proposes a method to build a statistical model using event logs for system failure prediction. To achieve the best prediction performance, prescreening and statistical variable

selection are adopted to select the best set of predictor events, coded as covariates in the statistical model. In-depth discussion of the prediction power of the model in terms of false alarm and misdetection probability is presented. Using a real-world example, the effectiveness of the proposed method is further confirmed.

Maikel Leemans at al paper presents hierarchy and recursion extension to the process tree model; and the first, recursion aware process model discovery technique that leverages hierarchical information in event logs, typically available for software systems [4]. The work can be positioned in-between reverse engineering and process mining. An implementation of the proposed approach is available as a ProM plugin. Experimental results based on real-life (software) event logs demonstrate the feasibility and usefulness of the approach and show the huge potential to speed up discovery by exploiting the available hierarchy.

Risto Vaarandi et al paper provides information on event logging which is supported by most applications, services, network devices, and other IT system components [5]. Well-known standards exist for event logging and widely used solutions have been developed for event log collection (such as sys logs, and Elastic Stack logs). Event logs contain valuable information about security incidents, but since large volumes of log data are generated in modern data centers and computer networks, the manual review of event logs is infeasible. In order to aid the human analyst, a number of data mining algorithms and tools have been proposed. Many suggested approaches are semi-automated, allowing for interactive discovery of event patterns from event logs. This knowledge can be used for various purposes like handling security incidents and developing event correlation rules. Here, the proposed algorithms assume that each line in the event log is a complete representation of some event. The algorithms divide the lines into clusters, so that lines from the same cluster are similar and matching the same line pattern. Instead of printing lines in each cluster, the algorithms output a line pattern for each cluster to the end user. Also, lines that do not fit into any of the detected clusters are arranged into a special cluster of outliers and reported individually. Due to their nature, clustering algorithms are able to identify not only event patterns that reflect regularities, but also unusual outlier events that deserve closer attention from security personnel.

Shi Shengyan et al paper emphasis on Information security behavior monitoring which is the basis of the information security management work [6]. Network information security behavior monitoring refer to collect a variety of network security events and network security log data from the safety equipment and information security systems and proceed data mining and conform the security situation and provide services for the later research of the information security and network trend.

In He Fang et al article, emphasis is on envision new artificial intelligence (AI)-enabled security provisioning approaches to overcome these issues while achieving fast authentication and

progressive authorization [7]. To be more specific, a lightweight intelligent authentication approach is developed by exploring machine learning at the base station to identify the prearranged access time sequences or frequency bands or codes used in IoT devices. Then we propose a holistic authentication and authorization approach, where online machine learning and trust management are adopted for achieving adaptive access control. These new AI-enabled approaches establish the connections between transceivers quickly and enhance security progressively so that communication latency can be reduced, and security risks are well controlled in large-scale IoT systems.

Jonghoon lee et al paper proposed technique converts multitude of collected security events to individual event profiles and uses a deep learning-based detection method for enhanced cyber-threat detection [8]. For this work, we developed an AI-SIEM system based on a combination of event proling for data preprocessing and different artificial neural network methods, including FCNN, CNN, and LSTM.

Anna L et al survey paper describes a focused literature survey of machine learning (ML) and data mining (DM) methods for cyber analytics in support of intrusion detection. Short tutorial descriptions of each ML/DM method are provided [9].

Hui Yang et al paper proposes a brain-like distributed control security (BLCS) architecture for F-RON in CPS by introducing a brain-like security (BLS) scheme [10]. BLCS can accomplish the secure cross-domain control among tripartite controllers verification in the scenario of decentralized F-RON for distributed computing and communications, which has no need to disclose the private information of each domain against cyber-attacks. BLS utilizes parts of information to perform control identification through a relation network and a deep learning of behavior library.

Tian Wang et al Paper explains about when malicious users or attackers simultaneously request the same physical nodes, it may lead to a failure of services as well as a security threat to the system [11]. In this article, the author has design a low-coupling system based on the edge computing platform to counter coupling problems. The edge computing platform acts as a middleware platform and provides the scheduling method. Based on the edge computing platform and artificial intelligence technology.

In the Hyunbum Kim et al article, author has introduced a new integrated virtual emotion system with AI applicability, called as IN-VEmoSYS, toward secure CPS-enabled smart cities [12]. The integrated virtual emotion system covers several subsystems such as virtual emotion barrier, virtual emotion map, and virtual emotion block. We describe their system settings, concepts, components, and operations, and also deal with AI applicability to those subsystems. Furthermore, we discuss future challenges and security issues that must be met to achieve secure advanced smart cities using the AI-enabled virtual emotion system.

Qiang Liu et al paper first provides details on Network slicing, which is a promising technology that allows network operators to create multiple virtual networks on top of a shared network infrastructure [13]. Later, it describes about various challenges to design secure network slicing solutions that can efficiently create end-to-end network slices for diverse cyber-physical systems. In this article, author discuss the challenges and security issues of network slicing, study learning-assisted network slicing solutions, and analyze their performance under the denial-of-service attack. Author also present a design and implementation of a small-scale test bed for evaluating the network slicing solutions.

3. Research Methodology:

Qualitative Samples and Quantitative Analysis methods will be used. In our study we will use either available standard data sets, or recorded data set from real use cases or will generate our own data set.

3.1 Problem Statement:

Cyberattack may happen anytime on any computer system connected over network. The attack may be done for data theft, identity theft, economic gains, revenge, or corporate espionage and for many more reasons.

Based on the attack category, the attacker may sometime attack specific machines or group of machines over the network. Post analysis, in the present scenario it is found that the system is used by attacker to target the services for the sake of commercial, ill legal activities, extortion and blackmail and other purposes.

Present available research is not enough to address the problems where the cyber security expert examines the huge stock of available logs to trace the attack events. Neither the mechanism is available to predict the attack scenarios from the available cyber-attack patterns. This is because log analysis is a time-consuming process and till date no solution is available to predict and analyze the logs in real time scenarios and also in the post analysis scenarios. Apart, it is not only logs, but several other factors are also needed to be examined to know the cyber-attack and its possible damage. In cyberattack scenarios, breach may occur due to various reasons like unattended open port, open gateways, unsecured domains, operating without SSL certificates, old hardware and software's, absence of hardening of network devices, absence of two factor authentications, relying on unsecured third-party services, allowing use of putty commands from back-doors and many more. These are few of the reasons which leads to cyberattacks.

Hence, it is expected to find out a novel method which can predict cyber security attack scenario in active or passive methods. One can use Artificial Intelligence (AI) based approach to predict such attacks where system can be trained to identify the cyber security attacks by analyzing various

system and network generated logs, authentication process, access to third-party services policy, previous case studies to create a system learning curve, etc. Research is needed to be done in order to solve the issues and problems in such Cyber Security Incident scenarios.

3.2 Methodology:

Samples/datasets will be collected and analyzed. The samples will be either collected from the live systems attack scenarios or the samples can be collected over the live network devices or if any available sample research data set is their then it will be used for testing purpose. But such dataset must be a standardize data set.

For the processing of the data both manual and automatic approach will be applied. In manual visual and tool-based analysis will be done. Further, the proposed machine learning approach will be used to study the samples and apply the proposed methodology to predict the attack scenarios and patterns. The results obtained from manual and machine learning approach will be compared. Further, they will be compared with previously available research outputs. The results obtained will be analyzed both qualitatively and quantitatively to maintain the accuracy and reliability of our findings.

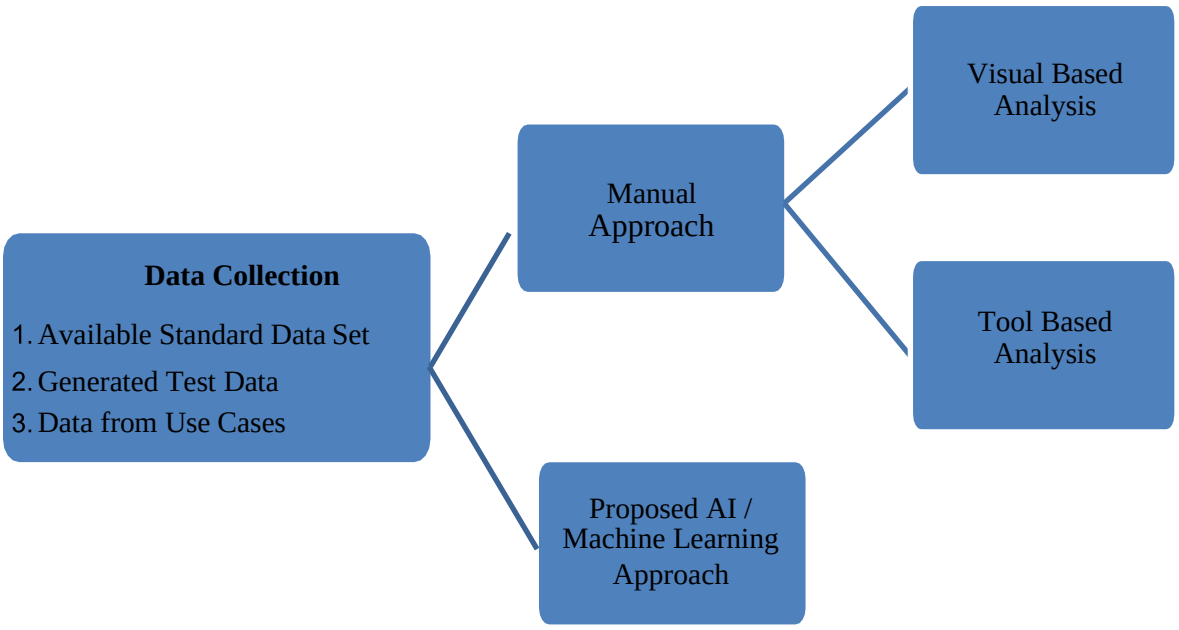

Fig 1: Proposed Research Work (Manual and Machine learning Based approach)

Data sources for our research may vary based on the cyberattack category. It could be scanning report for open ports, reports of vulnerability assessment, various types of logs, or even output generated after a command line instruction provided for assessment, etc. One must note that, even

in logs, their exist sub-categories, as shown below:

Hardware Source of Logs	Software Source of Logs
Individual Employee Systems (End-Point)	Operating System
Servers	Supporting Software Systems
Routers	Database Systems
Switches	Any sort of preventive services installed
Other Digital gadgets	Anti-virus systems, Firewalls, etc.

Internet Based Source of Logs
Internet Browsers
Plug-in
Etc.

These are major log components which are needed to be assessed to understand cyber security risk. Such log sources may be from Windows or Linux server logs, Firewall logs from Sonic wall, Juniper Network logs, Apache server logs, Active Directory Server logs, Anti-virus logs, Intrusion Detection System Logs, Intrusion Prevention System logs, MS SQL Server logs, Amazon Web Server logs, Hyper visors logs, Virtual Machine Server logs, Router and Switches logs, and many more such devices logs.

Hence, as research work progress and as per case-to-case basis, selection of input dataset will be decided.

Proposed AI Approach: At very first Data set will be selected based on cyberattack scenarios and then it will be normalized. System will accept such normalized dataset as input.

The proposed system will try to analyze and co-relate each of the incidents with the available dataset and based on the learning cycles it will initiate the prediction process. For every cycle of learning, the system will record the output against input and later will take it as a learning cycle. As system progress towards accommodating and executing a greater number of cybersecurity attack scenarios, the system gets trained, and this helps system to predict the cyberattack scenarios more precisely.

The generated prediction will be very hard to achieve as challenges faced by the proposed systems is that the provided input may come from varying systems (in different formats) and from different geographical regions. Hence, input may have different syntax and formats. Sometime cyberattack events may be sequential or parallel in occurrence or in other words, numerous machines in the network are attacked simultaneously or in sequential manner, and attack may span over various machines geographically located in different contents with different time zones,

logs may run in the size of terabytes, etc. and such attacks are done in parallel or serial attack methods. In all such scenarios, we propose to first normalize such database and fit in our proposed customized format so that system will be having streamlined input. Many more such challenges are associates with the development of Artificial Intelligence based Predictive system to analyze and detect the cyber-attack scenarios.

4. Objective:

a. Understanding the present cyberattack scenarios
b. Understanding the present cyberattack unearthing mechanism
c. Understanding the inputs used while unearthing the cyberattacks and its importance
d. To study and propose the AI/ML based approach for detecting the cyber security breach or attack scenarios through various inputs available from different sources.

5. Expected Outcomes:

With planned work undertaken, following expected outcomes are to be achieved.

a. Proposed research work will provide systematic approach to first analyze the available predictive methods of cyber security breach events or attacks.

b. It will aim to address and overcome the challenges and shortcoming in present solutions, which will help to analysis and to conclude the overall approach for cyber security and forensic investigations with accuracy.

c. It will also hold the technical and scientific validation for the judicial procedure.

d. Our research work is targeted and directed towards the present challenges and problems faced by cyber security and law enforcement agencies across the globe to analyze, identify and validate the cyber security breach events or attacks in real time scenarios.

e. Our research work is focused on solving various challenges currently faced by cyber security experts to analyze, co-related and predict the cyber security breach events or attacks.

References

1. Log Files Analysis for Network Intrusion Detection, Andre Brandao; Petia Georgieva, published in 2020 IEEE 10th International Conference on Intelligent Systems (IS), Date of Conference: 28-30 Aug. 2020, ISSN: 1541-1672.

2. Optimizing windows 10 logging to detect network security threats, Július Baráth, Published in: 2017 Communication and Information Technologies (KIT), Vysoke Tatry, Slovakia, Date of Conference: 4-6 Oct. 2017, INSPEC Accession Number: 17375055, ISBN:978-1-5386-2919-2.

3. Yuan, Shiyu Zhou, Crispian Sievenpiper, Kamal Mannar &Yibin Zheng, Event log modeling and analysis for system failure prediction, IIE Transaction, Volume 43, 2011 – Issue – 9, Published on 8 June 2011.

4. Maikel Leemans; Wil M. P. van der Aalst; Mark G. J. van den Brand, Recursion aware modeling and discovery for hierarchical software event log analysis, 2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER), 23-24 March 2018.

5. Risto Vaarandi; Markus Kont; Mauno Pihelgas, Event log analysis with the LogCluster tool, published in: MILCOM 2016 - 2016 IEEE Military Communications Conference, 1-3 Nov. 2016. INSPEC Accession Number: 16546511.Electronic ISSN: 2155-7586.

6. Shi Shengyan; Shen Xiaoliu; Zhao Jianbao; Ma Xinke, Research on System Logs Collection and Analysis Model of the Network and Information Security System by Using Multi-agent Technology, published in 2012 Fourth International Conference on Multimedia Information Networking and Security, 2-4 Nov. 2012. Print ISSN: 2162-8998.

7. He Fang, Angie Qi, Xianbin Wang, Fast Authentication and Progressive Authorization in Large-Scale IoT: How to Leverage AI for Security Enhancement, Published in: IEEE Network Journal (Volume: 34, Issue: 3, May/June 2020), INSPEC Accession Number: 19669267, Date of Publication: 02 June 2020, Print ISSN: 0890-8044.

8. Jonghoon lee, Jonghyun kim, Ikkyun kim and Kijun Han, Cyber Threat Detection Based on Artificial Neural Networks Using Event Profiles, Special Section on Artificial Intelligence In Cybersecurity, IEEE Access, Date of Publication - November 12, 2019, date of current version November 25, 2019.

9. Anna L. Buczak, and Erhan Guven, A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection, IEEE Communications Surveys & Tutorials, Vol. 18, No. 2, Second Quarter 2016, Date of Publication: October 26, 2015.

10. Hui Yang; Kaixuan Zhan; Michel Kadoch; Yongshen Liang; Mohamed Cheriet, BLCS: Brain-Like Distributed Control Security in Cyber Physical Systems, Date of Publication: 02 June 2020, Published in: IEEE Network Journal (Volume: 34, Issue: 3, May/June 2020), Print ISSN: 0890-8044, Electronic ISSN: 1558-156X.

11. Tian, Yuzhu Liang; Yi Yang; Guangquan Xu; Hao Peng; Anfeng Liu; Weijia Jia, An Intelligent Edge-Computing-Based Method to Counter Coupling Problems in Cyber-Physical Systems, Date of Publication: May/June 2020, Published in: IEEE Network (Volume: 34, Issue: 3, May/June 2020), Print ISSN: 0890-8044, Electronic ISSN: 1558-156X.

12. Hyunbum Kim; Jalel Ben-Othman, Toward Integrated Virtual Emotion System with AI Applicability for Secure CPS-Enabled Smart Cities: AI-Based Research Challenges and Security Issues, Date of Publication: 02 June 2020, Published in: IEEE Network Journal (Volume: 34, Issue: 3, May/June 2020), Print ISSN: 0890-8044, Electronic ISSN: 1558-156X.

13. Qiang Liu; Tao Han; Nirwan Ansari, Learning-Assisted Secure End-to-End Network Slicing for Cyber-Physical Systems, Date of Publication: 02 June 2020, Published in: IEEE Network (Volume: 34, Issue: 3, May/June 2020), Print ISSN: 0890-8044, Electronic ISSN: 1558-156X.

Research Scholar	**Research Guide**
(Shivaji Digambarrao Jadhav)	(Dr. Balaji R Bombade)

Date:

Place: Nanded.